PROCEEDINGS OF TDWG

Biodiversity Information Science and Standarts 1: e20066
doi: 10.3897/tdwgproceedings.1.20066

OPEN ACCESS

Conference Abstract

Developing a knowledge base on the habitats and reproductive conditions of Dipterocarps through information extraction

Roselyn Gabud[‡,§], Sandra Yap[|], Riza Batista-Navarro[¶,§], Sophia Ananiadou[¶]

‡ University of the Philippines Diliman, Quezon City, Philippines
§ University of the Philippines Los Baños, Laguna, Philippines
| Fareastern University, Manila, Philippines
¶ University of Manchester, Manchester, United Kingdom

Corresponding author: Roselyn Gabud (rsgabud@up.edu.ph)

Received: 06 Aug 2017 | Published: 07 Aug 2017

Citation: Gabud R, Yap S, Batista-Navarro R, Ananiadou S (2017) Developing a knowledge base on the habitats and reproductive conditions of Dipterocarps through information extraction. Proceedings of TDWG 1: e20066. https://doi.org/10.3897/tdwgproceedings.1.20066

Abstract

Dipterocarps, belonging to the family *Dipterocarpaceae*, are economically and ecologically important in the Philippines due to their timber value as well as contribution to wildlife habitat, climatic balance and stronghold on water releases. The supra-annual mass flowering of dipterocarps occurs in irregular intervals of two to ten years, possibly synchronously across Asia. Predicting the likelihood of their regeneration, to subsequently make plans regarding species for reforestation, can be aided by providing access to a knowledge base of dipterocarps, including information on the factors that affect their flowering and fruiting patterns. The content of the knowledge base could be enriched with literature-derived data on dipterocarp occurrence, reproductive condition and habitat.

We aim to develop information extraction methods to automatically extract concepts relevant to the reproductive cycle of dipterocarps to enable searching for more descriptive information of mass flowering from the literature. In previous work, we created a corpus of text selected from the Biodiversity Heritage Library (BHL), scholarly articles, books and government agency reports with manually labelled taxon names, geographic locations, dates, habitat descriptions, authorities, and names of herbaria (in the case of collected

specimens) to aid in determining the distribution of dipterocarps. Importantly, the species' reproductive condition, e.g., whether it is in fruit, in flower or sterile, was also annotated to enable the derivation of phenological patterns and the identification of factors that trigger mass flowering.

In this work, we focus our efforts on the automatic annotation of information on a species' reproductive condition, which we cast as a named entity recognition (NER) task. To this end, we have developed machine learning-based models building upon conditional random fields (Lafferty et al. 2001). The resulting new NER tool has been integrated as a new component in Argo (Rak et al. 2012) to allow for the linking of information on reproductive condition, with species names and habitat descriptions. This will eventually enable the generation of more descriptive occurrence data that includes information on reproductive and habitat conditions of dipterocarps. This serves as a step towards a more comprehensive basis of restoration efforts for dipterocarp forests.

Keywords

Philippine dipterocarps, Dipterocarpaceae, reproductive conditions, habitats, text mining, information extraction

Presenting author

Roselyn Gabud

Funding program

Newton Fund Institutional Links

Grant title

Conserving Philippine Biodiversity by Understanding Big Data (COPIOUS): Integration and analysis of heterogeneous information on Philippine biodiversity

References

- Lafferty J, McCallum A, Pereira FN (2001) Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. Proceedings of the Eighteenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc.

- Rak R, Rowley A, Black W, Ananiadou S (2012) Argo: an integrative, interactive, text mining-based workbench supporting curation. Database 2012: bas010. https://doi.org/10.1093/database/bas010